CHAPTER 1

The Road to Predictive Coding: Limitations on the Defensibility of Manual and Keyword Searching

Tracy D. Drynan and Jason R. Baron

Introduction

This chapter introduces the reader to the legal profession's interest in using information retrieval technology as a means to identify, filter, retrieve, and categorize data in an effort to respond to requests for discovery during litigation or investigations. In order to efficiently search through documents in the form of electronically stored information (ESI), lawyers need to be savvy in using human means as well as automated tools and techniques for reducing increasingly vast volumes of data and for identifying and categorizing relevant information. Maintaining the status quo of primary or sole reliance on manual review (i.e., document-by-document or linear review), coupled with only keyword searching, is increasingly seen as a flawed means for carrying out one's discovery obligations for at least a portion of the legal docket. Indeed, sole reliance on such methods in complex cases may in the near future raise questions about competence, including what should pass as due diligence in meeting one's professional obligations.

The modest aim of this chapter is to serve notice to the everyday practitioner of the issues surrounding still widely used manual and keyword search methodologies. A brief discussion of certain automated techniques used in conjunction with search methods is also included. We recognize that e-discovery practice has been transformed over the past decade and a half into a far more technical exercise, incorporating advanced search techniques borrowed from other disciplines. The subject of "information retrieval" is itself a science, and the interested reader will be able to explore in greater depth the mathematical, statistical, and algorithmic aspects of advanced search techniques in subsequent chapters in this volume.

Exhaustive Manual Review Is Increasingly an Outdated Approach

> Like the physical universe, the digital universe is large—by 2020 containing nearly as many digital bits as there are stars in the universe.[1]

Contemplating the number of stars in the universe is an amusing distraction, in the abstract; however, when considering the monumental task faced by legal professionals, where we are expected to make reasonable efforts to identify, aggregate, and categorize data to subsequently derive meaning out of ever increasing volumes of data, the celestial diversion loses its levity. The information retrieval task at hand, which currently necessitates a search through and analysis of very large data sets, increasingly poses all too real a burden.

[1] IDC iView, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things Executive Summary*, EMC^2 (Apr. 2014), http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm [hereinafter *Digital Universe*]. *See generally* John Foley, *Extreme Big Data: Beyond Zettabytes and Yottabytes*, Forbes (Oct. 9, 2013), http://www.forbes.com/sites/oracle/2013/10/09/extreme-big-data-beyond-zettabytes-and-yottabytes/.

The volume of global data is doubling in size every two years and is predicted to reach 44 zettabytes[2] by the year 2020.[3] Meaning, by 2020, there will be 5,200 gigabytes of data—equivalent to over 150,000 bankers boxes or 10 million pages of data—for every man, woman, and child on earth.[4] One would not conceive of analyzing 10 million printed pages, one by one, on a manual basis, as a practical or even possible method of identifying relevant and actionable information in any context, much less in a legal dispute with parties footing the bill.

This challenge only increases in magnitude with each passing year as data volume and variation continues to grow, while the cost of storing this data continues to decrease.[5] The result—current and certainly future volumes of data cannot humanly be reviewed, for litigation purposes, via a manual document-by-document process given the limited resources of labor, time, and the financial constraints of any given matter.[6] For this reason, leading e-discovery jurists

[2] A zettabyte is a very large volume of data equating to one trillion gigabytes; *see Zettabyte*, TechTerms, http://techterms.com/definition/zettabyte (last visited Aug. 1, 2016). *See also* Foley, *supra* note 1 ("In the hierarchy of big data, there are petabytes, exabytes, zettabytes, and yottabytes. After that, things get murky. The challenge is only partly one of coming to agreement on the right words to describe what lies beyond a yottabyte, which is septillion bytes.").

[3] *Digital Universe, supra* note 1.

[4] 1 gigabyte = 70,000 pages = 28 bankers boxes (each with 2,500 pages).
5,200 gigabytes = 150,000 bankers boxes = 10,500,000 pages.

[5] The Sedona Conference, *The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 15 Sedona Conf. J. 217, 228 (2014) [hereinafter *Sedona Search Commentary*] ("More recently, there has been a similar explosion in the use of instant and text messaging throughout organizations, including increasingly, through the use of mobile devices. In many organizations, the average worker maintains several gigabytes of stored data. At the same time, the costs of storage have plummeted from $20,000 per gigabyte in 1990 to less than 3.5¢ per gigabyte in 2013.").

[6] Bennett B. Borden, *The Demise of Linear Review*, DrinkerBiddle (Oct. 1, 2010), http://www.drinkerbiddle.com/resources/publications/2010/The-Demise-of-Linear-Review_2010 ("The explosive growth in the volume of data can create a crippling financial and administrative burden on parties responding to discovery requests to identify, collect, review, and produce data."); *see also* Harrison M. Brown, *Searching for an Answer: Defensible EDiscovery Search Techniques in the Absence of Judicial Voice*, 16 Chap. L. Rev. 407, 413 (2013) (citing Jason R. Baron & Michael D. Berman, *Designing a "Reasonable" E-Discovery Search: A Guide*

have questioned the continued viability of manual review as a sole means for identifying relevant information.[7] The Sedona Conference recognized this in its best practices commentary on search and retrieval:

> Particularly (but not exclusively) in large and complex litigation, where discovery is expected to encompass hundreds of thousands to hundreds of millions of potentially responsive electronic records, there is no reasonable possibility of marshalling the human labor required to undertake a document-by-document, manual review of the potential universe of discoverable materials.[8]

The era of "information inflation"[9] has brought into sharp relief the growing conflict between the principle of broad discovery,[10] supported by the current breadth of allowable discovery under the Federal Rules

for the Perplexed, in Managing E-Discovery and ESI: From Pre-litigation Through Trial 479, 481 (Berman et al. eds., 2011)).

[7] *See generally* U.S. v. O'Keefe, 537 F. Supp. 2d 14, 23–24 (D.D.C. 2008); Equity Analytics v. Lundin, 248 F.R.D. 331, 332–33 (D.D.C. 2008); Victor Stanley v. Creative Pipe, Inc., 250 F.R.D. 251, 260 (D. Md. 2008); Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on "Information Inflation" and Current Issues in E-Discovery Search*, 17 Rich. J.L. & Tech. 9 (2011), *available at* http://jolt.richmond.edu/v17i3/article9.pdf.

[8] *Sedona Search Commentary, supra* note 5, at 243 (Practice Point 1, "[i]n many settings involving electronically stored information, reliance *solely* on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary."). *See also* Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND Institute for Civil Justice 97, 99 (2012) (indicating that document review accounts for "$0.73 of every dollar spent on electronic production," and "computer categorized review strategy, such as predictive coding, [is] not only a cost-effective choice but perhaps the *only* reasonable way to handle many large-scale productions.") (emphasis in original).

[9] *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 Rich. J.L. & Tech. 10, ¶¶ 1–2 (2007), *available at* http://law.richmond.edu/jolt/v13i3/article10.pdf.

[10] Hickman v. Taylor, 329 U.S. 495, 507 (1947) ("Mutual knowledge of all relevant facts gathered by both parties is essential to proper litigation.").

of Civil Procedure (the Rules),[11] and the Rules' mandate to resolve disputes in a cost and time effective manner.[12] Broad—sometimes unduly broad—discovery requests for "relevant" information have led to the collection of expansive volumes of information, gradually rendering the "venerated process of 'eyes-only' manual review [as] no longer generally workable or economically feasible."[13] The volume and complexity of information, combined with the proven inaccuracy of manual review of this information, has driven the legal community, including clients, counsel, and jurists, to embrace alternative, hybrid methods of search, retrieval, and review of digital documents.[14]

This is not to say that automated technologies as a cost-efficient means to effectively and accurately identify relevant information will completely supplant the need for manual assessment of data. Manual review is still necessary to review seed sets of data used in supervised machine learning technologies, quality control processes, privilege review, and matters or investigations with smaller volumes of data.[15] It is also, a means to interrogate the data and ferret out the most

[11] Fed. R. Civ. P. 26(b)(1) (2016). *See infra* note 18 for a discussion of the 2015 Rules amendments.

[12] Fed. R. Civ. P. 1 (2016).

[13] *Sedona Search Commentary, supra* note 5, at 229.

[14] *See* Hon. Craig Shaffer, *Defensible by What Standard?*, 13 Sedona Conf. J. 212 (2012) (discussing the use of automated tools to review data in a more cost effective manner that is superior to manual review); *see also Sedona Search Commentary, supra* note 5, at 220 ("[J]ust as technology has given rise to these new litigation challenges, technology can help to solve them. The emergence of new discovery strategies, best practices, and processes, as well as new search and retrieval technologies are transforming the way lawyers litigate. Collectively, they provide opportunities for huge volumes of information to be reviewed faster, more accurately, and more affordably than ever before."); *id*. at 224 ("Alternative search tools may properly supplement simple keyword searching and Boolean search techniques. These include using various forms of computer- or technology-assisted review, machine learning, relevance ranking, and text mining tools which employ mathematical probabilities, as well as other techniques incorporating supervised and unsupervised document and content classifiers.").

[15] *See* Thomas Y. Allman, Jason R. Baron, & Maura R. Grossman, *Preservation, Search Technology, and Rulemaking*, 30 The Computer & Internet Lawyer 2 (2013); *see also Sedona Search Commentary, supra* note 5, at 244 ("Of course, the use of automated search methods is not intended to entirely eliminate the need for manual review; indeed, in many cases, both automated and manual searches will be conducted, with initial automated searches used for culling down a universe of material to

meaningful or actionable content, regardless of relevancy, or in light of an internal investigation. However, the remaining elements of manual review (and the skill sets that go with them) that apply in any of these contexts are now best used with, and, arguably empowered and magnified by, the latest available technologies.

Nevertheless, continued insistence on lawyers relying on historical "tried and true" methods in reviewing ESI creates a disjointed environment wherein the document review "is divorced from its primary purpose, to marshal the facts specific to a matter to prove a party's claims or defenses . . . ."[16] The would-be cost and labor of such a manual undertaking is anathema to the spirit and mandate of the Rules to resolve disputes in a "just, speedy, and inexpensive"[17] manner, as well as to the principle of proportionality now expressly identified in newly amended Rule 26.[18]

Long before the advent of the 'e' in legal discovery, information science, through the Blair and Maron study, debunked the myth that the human review of information has a level of accuracy that is acceptable or as high as imagined.[19] The 1985 study revealed that attorneys and supervising paralegals, employing an iterative process using search terms to identify relevant documents, achieved only a 20 percent

more manageable size (or prioritizing documents), followed by a secondary manual review process.").

[16] Bennett B. Borden, et al., *Why Document Review Is Broken*, William Mullens EDIG: E-Discovery and Information Governance 3 (May 2011), *available at* http://www.umiacs.umd.edu/~oard/desi4/papers/borden.pdf.

[17] Fed. R. Civ. P. 1 (2016).

[18] Amended Rule 26(b)(1), with effective date Dec. 1, 2015, states that "Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the amount in controversy, the importance of the issues at stake in the action, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable"; *see generally* The Sedona Conference, *The Sedona Conference® Best Practices Commentary on Proportionality in Electronic Discovery*, 14 Sedona Conf. J. 155 (2013) [hereinafter *Sedona Commentary on Proportionality*] (explaining the principle of proportionality).

[19] *See generally* David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Retrieval System*, 28 Communications of the ACM (1985).

level of accuracy, or recall, and, yet, assumed that they were achieving a 75 percent level of recall—a wide disparity given the historical assumption that manual/linear review of hardcopy and, subsequently, electronic information was reliable and accurate.[20] A similar analysis of the reliability of human manual or linear review was conducted in studies published as part of the TREC Legal Track (2006–2011).[21] As of at least 2011, scholarly research has demonstrated the ability of technology to assist with the identification and categorization of documents to a degree of accuracy that meets or surpasses the level achieved by human review—and doing so on a scale and speed that surpassed any human ability to make such determinations.[22]

In line with this research, more advanced search and review techniques, as discussed throughout this book, provide an opportunity to leverage limited resources, time, labor, and money, alongside advances in technology, to more accurately, affordably, and quickly identify data both relevant to and informative about a matter or investigation. However, notwithstanding such identifiable advantages, the use of the most advanced automated search techniques has not yet taken hold over discovery, even in clearly advantageous circumstances. Rather, the employment of keyword searches to cull a data set and the subsequent linear manual review of the results remain the most frequently used method of information retrieval.[23] The legal community at large, as well as their clients, appears to have an inherent distrust in newer technology and, yet, a misplaced faith in the accuracy and efficiency of the application of search terms (as well as the use of manual review).

[20] *See Sedona Search Commentary, supra* note 5, at 230 ("[T]here appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible—perhaps even perfect—and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual review of massive sets of electronic data sets (which it does not), the relative efficacy of that approach versus utilizing newly developed automated methods of review remains very much open to debate.").

[21] Reports of the findings of the TREC Legal Track (2006–2011) are available at http://trec-legal.umiacs.umd.edu/.

[22] *See* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 Rich J.L. & Tech. 11 (2011), *available at* http://jolt.richmond.edu/v17i3/article11.pdf. *See also* Maura R. Grossman & Gordon V. Cormack, *A Tour of Technology-Assisted Review*, Chapter 3 of this volume.

[23] *See Sedona Search Commentary, supra* note 5, at 229.

There is a fear that any method, other than a document-by-document, "eyes only" review, will fail to capture relevant and informative data.[24] Additionally, there are lingering concerns regarding the lack of scientific validity coupled with the "lack of knowledge or even confusion about the capabilities of automated search tools."[25] This mindset over the use of manual review to identify relevant data, and the belief that advanced information retrieval techniques lack scientific validity, is, in our view, wholly misplaced. This chapter, as well as the subsequent chapters in this volume, is aimed at addressing these misperceptions.

Overview of Electronic Tools and Process

Search and information retrieval, outside of selecting paper documents by hand, invariably involves using some form of technology. Even keyword search leverages the ability of machine programming to identify documents satisfying this type of search—however fallible. Before we discuss keyword searching proper, it may be appropriate to briefly review some of the leading automated means of reducing and organizing data sets, regardless of what automated search method is performed. An exigent need exists for lawyers to quickly and efficiently identify key information potentially relevant to a matter or investigation, beginning within the identification phase, and through to the final analysis and ultimate production phases of discovery.[26]

Identification of Data Sources

The identification of relevant data sources begins prior to the discovery phase, during the period of time in which preservation of data has been triggered.[27] This phase, more often than not, is a complicated

[24] *Id.*

[25] *Id.*

[26] *See generally* Electronic Discovery Reference Model, *available at* http://www.edrm.net/resources/guides/edrm-framework-guides) (last visited Aug. 1, 2016) [hereinafter EDRM]. For an alternative schematic workflow worth considering, *see* Ralph C. Losey, www.edbp.com (Electronic Discovery Best Practices Model) (last visited Aug. 1, 2016).

[27] *See generally* Paul W. Grimm, Michael D. Berman, Conor R. Crowley, & Leslie Wharton, *Proportionality in the Post-Hoc Analysis of Pre-Litigation Preservation Decisions*, 37 U. Balt. L. Rev. 394 (2008).